MATLAB TUTORIAL 1

Figure: Old Faithful erupting

MATLAB TUTORIAL 1

Summary statistics for waiting time

mean 72.4 mins

median 76 mins

mode 78 mins

standard deviation 13.7 mins

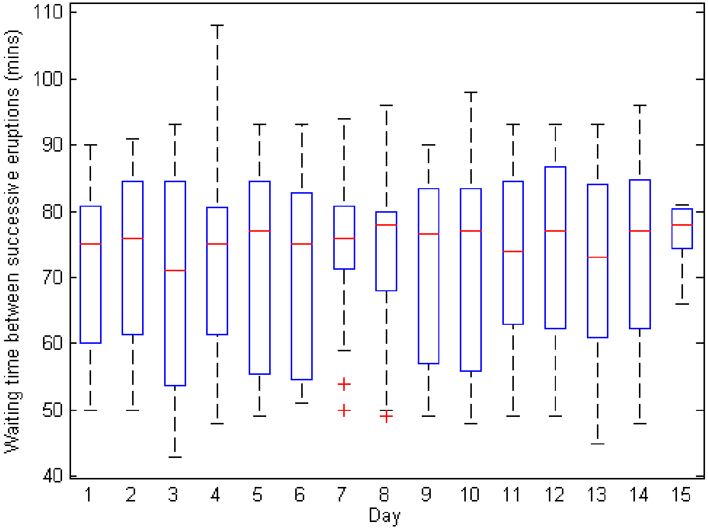
Q1 60.0 mins

Q3 82.3 mins

MATLAB TUTORIAL 1: BOX PLOTS

```
 % boxplot by day
figure()
boxplot(waiting, day)
xlabel('Day')
ylabel('Waiting time between successive eruptions
(mins)')
```

MATLAB TUTORIAL 1



8. How are the ends of the whiskers determined? Can you change this?
   `'whisker' Maximum whisker length W. Default is W=1.5.`
9. What patterns in waiting time, if any, do you notice? What can you say about day-to-day variation? How long do you predict you need to wait for the next eruption?
   Most waiting times fall in the interval [50, 100], with some exceeding this range. The variation is pretty much constant on most days with significant exceptions for Days 4 (too large) or Day 15 (too low). Some outliers observed for Days 7 & 8 too. Median time is pretty much the same each day, at around 75 mins, so that would be a sensible prediction.

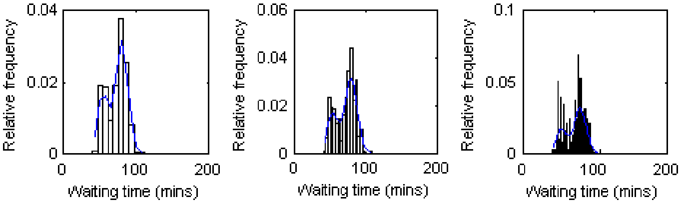
MATLAB TUTORIAL 1: HISTOGRAMS

```
 figure()
nbins = [ 10 20 50 ]
for bi = 1 :  3
bins = linspace(min(waiting), max(waiting),
nbins(bi));
freq = hist(waiting, bins); *
class = bins(2) - bins(1); *
relfreq = freq/(sum(freq)*class); *
ksestimate = ksdensity(waiting, bins);
subplot(1, 3, bi);
hold on;
bar(bins, relfreq, 1, 'FaceColor', 'w');
plot(bins, ksestimate, 'b-');
end
hold off
```

Or use the `histogram` command with option `'Normalization'` `'pdf'`.

This calculates and plots the relative frequency histogram and avoids using `hist`, `bar`, and calculating the relative frequency by hand.

MATLAB TUTORIAL 1

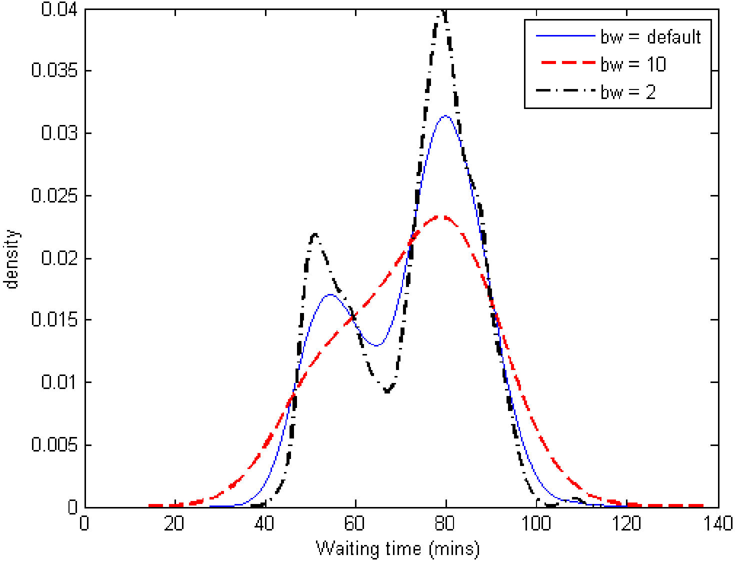


9. How many bins would you choose to best represent the distribution of waiting times? A commonly used criterion to determine the number of bins is Sturges' formula: $1 + \log_2 n$, where n is the sample size - do you think this is a good choice for these data?
   Sturges' formula gives 9.15 as starting point for the number of classes. More appropriate for normal data (also works better for large samples).
10. Comment on the histogram and kernel density smooth versus the boxplot for the "Old Faithful" waiting times.
   The boxplot does not give enough information on mode or the shape of the distribution of the waiting times, whereas the histogram and kernel density plot do. Boxplots can be used for getting the median, the quartiles, as well as whether there exist outliers. Skewness can be observed using all three plots (assuming the right number of bins/bandwidth is chosen).

MATLAB TUTORIAL 1: KERNEL DENSITY

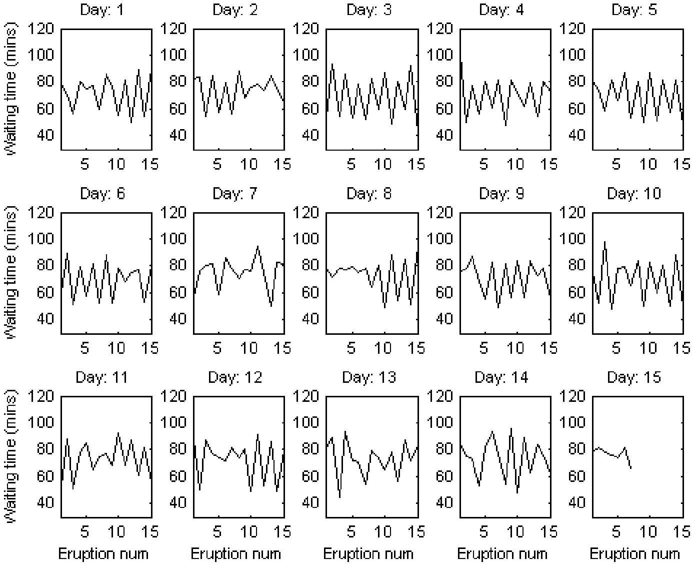
```
 [f,xi,bw] = ksdensity(waiting)
figure()
plot(xi, f)
hold on
(f,xi] = ksdensity(waiting,'width',10);
plot(xi,f,'--r','LineWidth',1.5);
(f,xi] = ksdensity(waiting,'width',2);
plot(xi,f,'-.k','LineWidth',1.5);
legend('bw = default','bw = 10','bw = 2');
hold off
```

MATLAB TUTORIAL 1: KERNEL DENSITY



```
for mi = 1 :  15
loc = find( day == mi );
subplot(3, 5, mi);
hold on;
plot(waiting(loc), 'k-');% line plot
set(gca, 'Box', 'On', 'FontSize', 8);
xlim([1 15]);
ylim([30 120]);
if mod(mi, 5) == 1
ylabel('Waiting time (mins)'); % label y-axes
end
if mi > 2*5
xlabel('Eruption num'); % label x-axes
end
title(['Day:  '  num2str(mi)]);
end
```

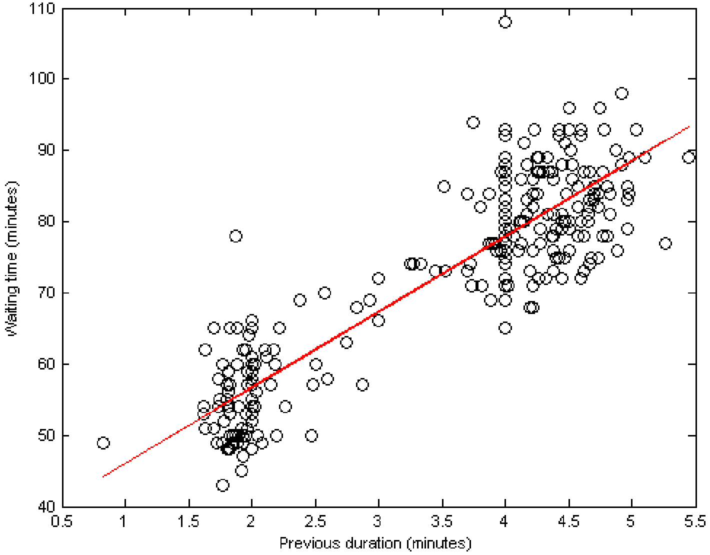
MATLAB TUTORIAL 1: TIME SERIES



MATLAB TUTORIAL 1: SCATTER PLOT

```
n = length(duration);
% do linear regression
lagduration = lagmatrix(duration, 1)
Y = waiting;
X = [ ones(n, 1) lagduration ];
B = regress(Y, X);
% fit linear regression (predicted values)
waitingest = B(1) + B(2)*lagduration;
% plot data and line of best fit
figure(4); clf;
hold on;
plot(lagduration, waiting, 'ko'); % plot data
plot(lagduration, waitingest, 'r-'); % plot line of
best fit
```

MATLAB TUTORIAL 1: SCATTER PLOT



What does this plot suggest to you? Do you think the linear fit is appropriate? What is your predicted waiting time now? Is there other information you could use to improve your prediction?
There is not a linear pattern in the data - waiting time can not be predicted linearly using the previous duration. Perhaps a non-linear model (e.g. polynomial regression) would be more appropriate to campture the relationship between the two. Based on the duration of the last eruption (4 minutes), the linear model would suggest a waiting time of 70 minutes. If we had more variables in the data set (e.g. weather conditions), we could have used these as covariates to get an improved prediction.

MATLAB TUTORIAL 1: RAW DATA

waiting	duration	day
80	4.02	1
71	2.15	1
57	4	1
80	4	1
75	4	1
77	2	1
60	4.38	1
86	4.28	1
77	2.03	1
56	4.83	1
81	1.83	1
50	5.45	1
89	1.62	1
54	4.87	1
90	4.38	1
73	1.77	1
60	4.67	1
83	2	1
65	4.73	1

MATLAB TUTORIAL 1: PREDICTIONS

Looking at the raw data, some of the durations are recorded as integer values, with less precision than others - why do you think this is?

MATLAB TUTORIAL 1: PREDICTIONS

Looking at the raw data, some of the durations are recorded as integer values, with less precision than others - why do you think this is? These were night time durations.

MATLAB TUTORIAL 1: PREDICTIONS

Looking at the raw data, some of the durations are recorded as integer values, with less precision than others - why do you think this is? These were night time durations. Possibly the night time data were estimated not measured?

MATLAB TUTORIAL 1: PREDICTIONS

Looking at the raw data, some of the durations are recorded as integer values, with less precision than others - why do you think this is? These were night time durations. Possibly the night time data were estimated not measured? How might this affect your predictions - how might you deal with this?

MATLAB TUTORIAL 1: PREDICTIONS

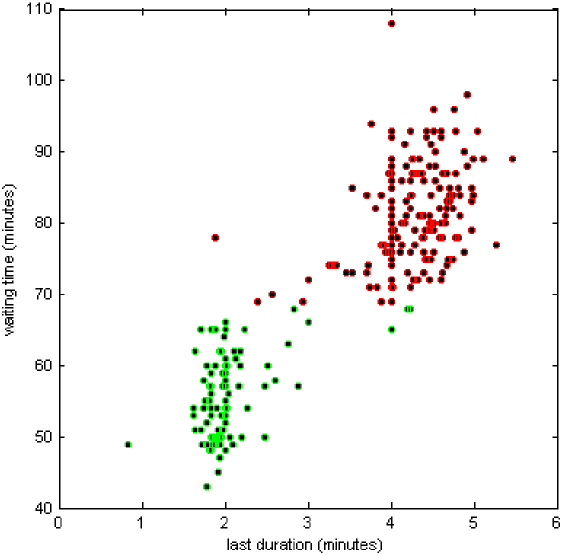
Looking at the raw data, some of the durations are recorded as integer values, with less precision than others - why do you think this is? These were night time durations. Possibly the night time data were estimated not measured? How might this affect your predictions - how might you deal with this?

Having less precision in some of the measurements can potentially lead to false conclusions drawn as predictions are made according to a model coming from imprecise data. A way to deal with that would be removing these data points which have been estimated and fitting a model using only accurate (i.e. daytime) measurements.

MATLAB TUTORIAL 1: CLUSTER

```
X = [ lagduration waiting ];
K = 2;
C = kmeans(X, K);
col1 = 'r'; col2 = 'g'; col3 = 'b'; col4 = 'm';
hold on
plot(lagduration, waiting, 'k.')
for c=1:K
loc = find( C == c );
plot(lagduration(loc), waiting(loc), 'color', colc,
'marker', 'o', 'markersize', 4, 'LineStyle', 'None');
end
```

MATLAB TUTORIAL 1: CLUSTER



MATLAB TUTORIAL 1: QUESTIONS

1. Compare the information revealed by each graph. What is gained (or lost) by each representation?
2. How useful were the summaries of location for predicting the expected waiting time to the next eruption?
3. What is your final prediction of waiting time? Can you predict more than one waiting time ahead? Will your predictions apply in 2016, and why/why not? Which graphical representation would best communicate your predictions? What other information would you provide?
4. You decide to stay in the park for the rest of the day to collect some more data and validate your prediction. The data are in `faithful15`. Import `faithful15` into MATLAB. How close are your final predictions for the next waiting times in comparison with the actual values?
5. What have you learned about (a) variation and (b) making predictions?

MATLAB TUTORIAL 1: ANSWERS

1. Boxplots give important information about the median, quartiles, skewness, variation. Histograms and kernel density plots give information on the shape of a distribution. Scatter plots display the data in such a way that any linear relationships between variables can be identified and predictions can be made. Time series plots display seasonal variation and can reveal trends/patterns.
2. They can be rather useful but just for a vague and 'safe' prediction (i.e. not extreme); in fact, visualising the data can help us get a better understanding of what the data really looks like and this way we can make better predictions.

MATLAB TUTORIAL 1: ANSWERS

3. Predicting more than one waiting time ahead is not easy but one can instead use the waiting time between every second/third etc. eruption as a target variable. The predictions are not necessarily going to apply in 2016 as we are extrapolating and we would need to assume that the waiting time only depends on the previous duration and on nothing else (which need not hold). The graphical representation that would be most appropriate is a scatter plot with a fitted line/curve, while time series plots could also be used as additional evidence (in the case of a presence of any trend).
4. According to the linear model, we are 84-70=14 minutes off the true waiting time. If we blindly went for the median from the boxplots, we would be 84-75=9 minutes off while going for the mode at around 81 minutes from the kernel density plots, we are 84-81=3 minutes off (notice that this does not mean we should go for the latter in general).